PHYSICAL REVIEW B VOLUME 41, NUMBER 2 15 JANUARY 1990-I

Role of tip electronic structure in scanning tunneling microscope images

J. Tersoff
IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York 10598
(Received 14 November 1989)

Tunneling to a spherical metal tip is analyzed, going beyond the *s*-wave approximation for the tip wave functions. For metal surfaces, the scanning tunneling microscope image is found to correspond to a contour of constant surface local density of states, under rather general assumptions. However, for semimetallic or semiconducting surfaces, the image may deviate in a simple but crucial way from the local density of states, as illustrated by the case of graphite.

A major obstacle to the quantitative understanding of scanning tunneling microscopy[1] (STM) is the uncertainty regarding the structure of the tip, and the resulting arbitrariness in the treatment of the tip in all theoretical studies. One theory, the "*s*-wave tip" model of Tersoff and Hamann,[2] is of particular interest despite its extreme simplicity, because it leads to a direct interpretation of the STM image as a contour of constant surface local density of states (LDOS).

By considering a more general picture of the tip, this paper both extends that interpretation, by showing that in most cases it remains valid under more general assumptions than those made in Ref. 2; and modifies the interpretation, by identifying the cases where it breaks down. These results represent the most precise statement to date of the connection between the STM image and the surface LDOS.

Specifically, the tip electronic structure is treated in two simple limits, which correspond roughly to small and large Fermi wave vector k_F. In the case of a spherical tip, the STM image can be calculated for both limits. By comparing these, we see to what extent the STM image depends on the tip electronic structure.

One finds that for metal surfaces, these different limits lead to images essentially identical to each other and to the surface LDOS. It is therefore reasonable to conclude that any realistic tip electronic structure, with an intermediate k_F, will also lead to an image which corresponds directly to the LDOS.

However, for some semimetallic or semiconducting surfaces, the lowest Fourier component of the image may deviate drastically from the LDOS, depending sensitively on the tip electronic structure. Nevertheless, the higher Fourier components, which generally carry the useful information, still closely follow the LDOS. Thus we conclude that the correspondence between the STM image and the surface LDOS is quite general, except *in certain cases* for the lowest Fourier component of the image. These cases are defined below.

The derivation of these results proceeds as follows. In perturbation theory, the tunneling current is simply[2]

$$J = \frac{2\pi e}{\hbar} \sum_{\mu,\nu} f_{\mu\nu} |M_{\mu\nu}|^2 , \tag{1}$$

where $M_{\mu\nu}$ is the tunneling matrix element between states ψ_μ and ψ_ν of the surface and tip. Here $f_{\mu\nu}$ is a shorthand for the terms representing Fermi occupancies and energy conservation, and implicitly depends on voltage, i.e.,

$$f_{\mu\nu} = \{f(E_\mu)[1-f(E_\nu)] - f(E_\nu)[1-f(E_\mu)]\} \times \delta(E_\mu + V - E_\nu) , \tag{2}$$

where $f(E)$ is the Fermi function, V is the applied voltage (in units of energy, i.e., eV), and E_μ is the energy of ψ_μ relative to the Fermi level of that electrode.

Numerous specific models of the tip have been used in theoretical treatments of STM.[2-8] Here the tip wave function is expanded in the form

$$\psi_\mu = \sum_i a_{i\mu} \phi_i , \tag{3}$$

where ψ_μ is a particular tip wave function with energy E_μ, and $\phi_i(\mathbf{r}) = \exp(-\kappa|\mathbf{r}-\boldsymbol{\tau}_i|)/r$ is an *s*-wave function centered at the position $\boldsymbol{\tau}_i$ of atom i, with $\kappa = \hbar^{-1}[2m(V - E_\mu)]^{1/2}$. (Throughout this discussion, normalization and other constant prefactors are omitted for simplicity wherever possible.)

Using the result of Ref. 2 for tunneling to an *s*-wave function, the tunneling matrix element is then

$$M_{\mu\nu} \propto \sum_i a_{i\mu}^* \psi_\nu(\boldsymbol{\tau}_i) , \tag{4}$$

and the current (1) becomes

$$J \propto \sum_{\mu,\nu} f_{\mu\nu} \sum_{i,j} \psi_\nu(\boldsymbol{\tau}_i) \psi_\nu^*(\boldsymbol{\tau}_j) a_{i\mu}^* a_{j\mu} . \tag{5}$$

At this point, one could model a specific tip geometry and electronic structure by appropriate choice of $\boldsymbol{\tau}_i$ and $a_{i\mu}$. By including sites separated by arbitrarily small distances, one could, moreover, model non-*s*-wave atoms. Specific choices can lead to interesting effects, such as enhanced resolution[7] or distorted images.[8]

However, the primary concern here is in developing *generic* models. Very specific tip models seem more appropriate in discussing particular experiments, where the tip has been characterized either directly[9] or by inference,[8] than in a general theoretical treatment.

The role of the atom positions $\boldsymbol{\tau}_i$ in defining the geometric structure of the tip is quite transparent. However, the role of the coefficients $a_{i\mu}$ in modeling the tip electronic structure is more problematic. Here we consider two different assumptions for $a_{i\mu}$. If the Fermi wave vector k_F of the tip is small, so that for a relevant tip

length scale[10] d, $k_F d \ll 1$, it is reasonable to consider a model where all atoms contribute coherently for a given tip wave function ψ_μ, i.e., $a_{i\mu} = \alpha_\mu A_i$. In that case the current (5) becomes

$$J^{\text{coherent}} \propto \sum_\nu D_\nu \left| \sum_i A_i^* \psi_\nu(\tau_i) \right|^2 , \tag{6}$$

where $D_\nu = \sum_\mu f_{\mu\nu} |\alpha_\mu|^2$ is closely related to the total density of available tip states at energy $E_\mu = E_\nu - V$.

On the other hand, for $k_F d \gg 1$, it is natural to assume that the relative phases of the coefficients $a_{i\mu}$ will be essentially random from atom to atom. In that case $\sum_\mu a_{i\mu}^* a_{j\mu} \rightarrow \delta_{ij} \sum_\mu |a_{i\mu}|^2$, giving

$$J^{\text{incoherent}} \propto \sum_{i,\nu} D_{i\nu} |\psi_\nu(\tau_i)|^2 , \tag{7}$$

where $D_{i\nu} = \sum_\mu f_{\mu\nu} |a_{i\mu}|^2$ reflects the density of available tip states at energy $E_\nu - V$ on atom i.

For typical metals, $k_F d$ may be of order unity, so neither extreme corresponds strictly to reality. The point here is simply to motivate two limits for treating the difficult part of the problem, the wave-function coherence within the tip. In this way, bounds are established for the true behavior, since we expect the image to vary monotonically with k_F of the tip.

Equations (6) and (7) could be used directly for modeling possible tip structures. However, instead, in order to make a more "generic" model, the discrete atoms are replaced here with an integral over a continuum of atom positions, uniformly distributed over a surface S. To be consistent with the assumption of a uniform distribution, we take $D_{i\nu}$ and A_i as independent of the site i. (Without loss of generality we take $A_i = 1$.) Then the notation D_ν can be used for $D_{i\nu}$ as well, since the two are now equivalent.

The current (6) from ψ_ν is then

$$J^{\text{coherent}} \propto \sum_\nu D_\nu \left| \int_S \psi_\nu(\mathbf{r}) d\mathbf{r} \right|^2 , \tag{8}$$

where the integral is over the surface S, which implicitly depends on tip position; or from (7),

$$J^{\text{incoherent}} \propto \sum_\nu D_\nu \int_S |\psi_\nu(\mathbf{r})|^2 d\mathbf{r} . \tag{9}$$

The term $\sum_\nu D_\nu$ is simply the usual integration[11] over the energy range (determined by the voltage) which contributes to the tunneling, weighted by the tip density of states. For small voltage, this reduces to projecting out those ψ_ν at the Fermi level, as in Ref. 2. For many purposes, it is adequate to take D_ν as constant over the allowed energy range.

Equation (9) represents a particularly intuitive and appealing result. It says that, *in the incoherent limit*, the total tunneling current is simply proportional to the surface LDOS integrated over the tip; so the image is the LDOS *convoluted with the shape of the tip.* This result has been implicitly *assumed* in some previous discussions of the role of tip shape; but here the result is derived, along with the condition for its validity.

To proceed further, and evaluate the integrals (8) and (9), we need explicit forms for the wave functions ψ_ν and the tip shape S. As usual, ψ_ν can be expanded in the form[2]

$$\psi_\nu(\mathbf{r}) = \sum_Q b_Q \exp(i\mathbf{Q}\cdot\mathbf{x}) \exp(-\kappa_Q z) , \tag{10}$$

where $\kappa_Q = (\kappa^2 + Q^2)^{1/2}$, $\mathbf{Q} = \mathbf{k}_\parallel + \mathbf{G}$, $\mathbf{k}_\parallel$ is the two-dimensional wave vector, and $\mathbf{G}$ is a two-dimensional reciprocal-lattice vector of the surface.

For simplicity, the tip is taken as spherical, with the radius of curvature R. The resulting current does not differ appreciably from that for a parabola or other smooth shape with the same curvature, as long as $\kappa R \gg 1$. This condition is well satisfied except when R is so small as to correspond to a single atom, in which case the spherical model is particularly appropriate.

The integral in (8) can be evaluated much as in Ref. 2, giving[12]

$$J^{\text{coherent}} \propto R^2 \sinh^2(\kappa R) \sum_\nu D_\nu |\psi_\nu(\mathbf{r}_t)|^2 , \tag{11}$$

where $\mathbf{r}_t$ is the position of the *center of curvature* of the tip. This is essentially equivalent[12] to the earlier result of Tersoff and Hamann.[2] The sum over ν is simply the surface LDOS at $\mathbf{r}_t$, multiplied by the tip state density and integrated from $E_F - V$ to E_F.

The treatment of the incoherent case (9) is trickier, because an explicit form is needed for $|\psi_\nu(\mathbf{r})|^2$. From (10),

$$|\psi_\nu(\mathbf{r})|^2 = \sum_{Q,Q'} b_Q b_{Q'}^* \exp[i(\mathbf{Q}-\mathbf{Q}')\cdot\mathbf{x}] \times \exp[-(\kappa_Q + \kappa_{Q'})z] . \tag{12}$$

For any given Fourier component $\mathbf{G}$ of the charge, the term in (12) which (if it exists) decays most slowly with distance is that with $\mathbf{Q} = -\mathbf{Q}' = G/2$.

If states at both the center and edge of the surface Brillouin zone (SBZ) fall near enough to E_F to contribute current, so that $\mathbf{Q} = -\mathbf{Q}' = G/2$ is always satisfied for some state, then asymptotically

$$|\psi_\nu(\mathbf{r})|^2 \simeq \sum_G c_G \exp(i\mathbf{G}\cdot\mathbf{x}) \times \exp[-(4\kappa^2 + G^2)^{1/2} z] . \tag{13}$$

Substituting (13) into (9) gives, for a spherical tip,

$$J^{\text{incoherent}} \propto R \sinh(2\kappa R) \sum_\nu D_\nu |\psi_\nu(\mathbf{r}_t)|^2 . \tag{14}$$

In this case the coefficient is linear in R, consistent with the macroscopic limit originally considered by Binnig *et al.*[1]

However, the important dependence, that on $|\psi_\nu(\mathbf{r}_t)|^2$, is identical to the coherent case (11). Thus the proportionality between tunneling current and $\sum |\psi_\nu|^2$ (i.e., LDOS) is apparently quite general.

[Actually, the conditions for the *approximate* validity of (13) turn out to be considerably less stringent than might be expected. Equation (13) typically applies (with a slightly renormalized value of κ) even at distances which cannot be considered asymptotic, because the more rapidly decaying terms scale with the asymptotically dominant one. Moreover, the requirement of having states with $k_\parallel = 0$ and with $k_\parallel = G/2$ is far from strict. In fact, a state with $k_\parallel = G/4$ is close enough to both the center and

edge to make (13) accurate enough for most purposes. Thus (13) is a good approximation for essentially any metal. The interested reader is referred to Ref. 13 for further discussion.]

The important point here is that, insofar as (13) applies, *the two opposite approximations give essentially identical images.* The same would presumably be true for a realistic intermediate degree of coherence as well. Thus the result of Tersoff and Hamann,[2] that the current is proportional to the local density of states at the center of curvature of the tip, is *for metals* [where (13) is rather accurate], much more generally valid than the original derivation might suggest.

For semiconducting surfaces at low tunneling voltage, the situation is a bit different. Tunneling then occurs to states at the band edge, so only a small pocket of the Brillouin zone contributes. Then for a given Fourier component $\mathbf{G}_c$ of the charge, there may be no $\mathbf{G}$ such that $\mathbf{k}_\parallel+\mathbf{G}=\mathbf{G}_c/2$ is well satisfied for the available small range of $k_\parallel$.

However, it turns out[13] that the relative errors in (13) are still not large *except* for the lowest Fourier component $\mathbf{g}$ of the image. The image is determined by the ratio of the higher Fourier components of the charge to the $G=0$ component. This ratio in turn depends sensitively on the difference in decay constants. For higher Fourier components, $k_\parallel \ll G$, so the details of the electronic structure are relatively unimportant, just as for a metal. Even for the second-lowest Fourier component of the image, (13) is reasonably well obeyed. However, the difference in decay constants between the $G=0$ and g Fourier components, which determines the g component of the image, can vary drastically, from zero to about twice the metallic value, if $k_\parallel$ is restricted to the edge or center of the SBZ.

For the $\mathbf{g}$ Fourier component, the correspondence between the image and the LDOS may therefore break down completely. The most extreme and most interesting example of this breakdown occurs when only states at the edge of the surface Brillouin zone contribute. This case has already been discussed elsewhere, because it also leads to anomalous corrugations[14] and enhanced resolution.[13]

Consider the limit where only a single wave function contributes to the tunneling current, and that wave function falls exactly at the edge of the surface Brillouin zone. To illustrate the resulting dependence of the image on the tip, it is sufficient to consider a two-plane-wave model, which in any case accurately describes the wave function at large distances.[14] Then

$$\psi \propto \cos(gx/2)\exp[-\kappa^2+g^2/4)^{1/2}z]\,. \tag{15}$$

The image for the incoherent tip must be calculated directly from (9) and (15), since (13) does not apply in this case. One obtains

$$z=\{\ln[1+\alpha_R\cos(gx)]-\ln C\}/(4\kappa^2+g^2)^{1/2}\,, \tag{16}$$

$$\alpha_R=(1+g^2/4\kappa^2)^{1/2}\sinh(2\kappa R)/\sinh[2(\kappa^2+g^2/4)^{1/2}R]\,.$$

(C is proportional to the tunneling current, and in this special case only shifts the image uniformly in z.) For $R\rightarrow 0$, this reduces to the same image as in the coherent case,[14] shown in Fig. 1, with a periodic array of singular dips associated with nodes in the wave function.

For finite R, however, the singularity in the image disappears, and the image becomes progressively smoother with increasing R. In fact, in the incoherent limit and for $R \gg 1/2\kappa$, the tip radius has exactly the same effect here as for a metal, in suppressing nonzero Fourier components of the image. This is true even though, unlike the metallic case, increasing the tip height has no such effect on the lowest nonzero component.

It is worth considering a specific example to illustrate the magnitude of these effects. The case of graphite has commanded considerable attention, and may be treated using (15), with appropriate parameters, for a path connecting hollow sites.[14] (Such a path passes over the bonds but not the atoms, and so does not sample quite the full corrugation.) An elegant treatment of this case for a somewhat different tip, but also not assuming s-wave wave functions, has been reported by Sacks *et al.*[15]

Figure 1 shows the theoretical graphite image for the coherent model, and for the incoherent model with two values of R. For comparison, clean UHV experiments have seen corrugations of about 0.3 Å or larger, although there is still considerable uncertainty.[16]

For a one-atom tip, a reasonable value of R would be the metallic radius for W or Ir, about 1.4 Å. The resulting corrugation in Fig. 1 is slightly under 0.2 Å, somewhat smaller than that seen experimentally.[16] Of course, just as the coherent limit neglects the smoothing due to finite R, the incoherent limit exaggerates the smoothing by an amount which is difficult to estimate. The most important point is that in a case such as graphite, unlike the case of metals, the two limits give radically different results, and the real image is thus quite different from the LDOS (i.e., from the image in the s-wave model or the $R\rightarrow 0$ limit).

In an *ad hoc* attempt to model a realistic case intermediate between incoherent with $R=1.4$ Å, and coherent (i.e, incoherent with $R=0$), we consider the incoherent limit with $R=0.7$ Å. This gives a corrugation of over 0.4 Å, also shown in Fig. 1, which is apparently consistent with experiment.[16] Thus for graphite, unlike metal surfaces, a precise knowledge of the tip electronic structure is crucial for a quantitative analysis.

In conclusion, we have considered two simple models for the STM tip. By comparing these, we address the

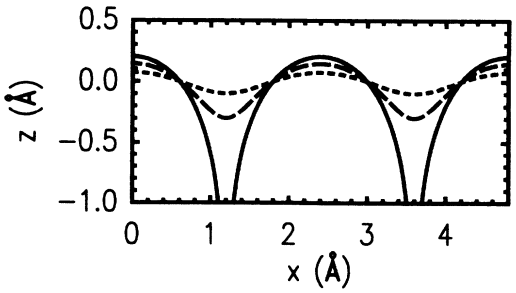

FIG. 1. Calculated STM image from Eq. (16), with parameters appropriate for graphite, for $R\rightarrow 0$ (equivalent to coherent model) (solid line), $R=0.7$ Å (dashed line), and $R=1.4$ Å (dotted line).

question of to what extent the STM image depends on the tip electronic structure. For metals, the earlier conclusion[2] that STM measures the surface LDOS is shown to be valid under much more general assumptions than those of Ref. 2, i.e., an s-wave tip is not required. However, in tunneling to semiconducting or semimetallic surfaces at low voltage, where only a small pocket of the surface Brillouin zone contributes, the actual STM image may deviate drastically from the results of the s-wave tip model, and from the LDOS. Fortunately, this deviation is confined to the lowest Fourier component of the image. The simple s-wave tip model may therefore be used even in this case to interpret data, as long as its one shortcoming is borne in mind.

I am grateful to A. Baratoff for discussions, and to N. D. Lang for comments on the manuscript.

[1]G. Binnig, H. Rohrer, Ch. Gerber, and E. Weibel, Phys. Rev. Lett. **49**, 57 (1982). For reviews, see G. Binnig and H. Rohrer, Rev. Mod. Phys. **59**, 615 (1987); P. K. Hansma and J. Tersoff, J. Appl. Phys. **61**, R1 (1987).

[2]J. Tersoff and D. R. Hamann, Phys. Rev. B **31**, 805 (1985); Phys. Rev. Lett. **50**, 1998 (1983).

[3]N. Garcia, C. Ocal, and F. Flores, Phys. Rev. Lett. **50**, 2002 (1983).

[4]E. Stoll, A. Baratoff, A. Selloni, and P. Carnevali, J. Phys. C **17**, 3073 (1984).

[5]N. D. Lang, Phys. Rev. Lett. **56**, 1164 (1986); **55**, 230 (1985).

[6]W. Sacks, S. Gauthier, S. Rousset, and J. Klein, Phys. Rev. B **37**, 4489 (1988).

[7]C. J. Chen, J. Vac. Sci. Technol. A **6**, 319 (1988); and (unpublished).

[8]H. A. Mizes, S. Park, and W. A. Harrison, Phys. Rev. B **36**, 4491 (1987).

[9]Y. Kuk, P. J. Silverman, and H. Q. Nguyen, J. Vac. Sci. Technol. A **6**, 524 (1988).

[10]It is tempting to take d as an atomic spacing, but it is perhaps more appropriate here to take it as the size of the tip region from which appreciable current flows, $d \sim (R/\kappa)^{1/2}$.

[11]J. Tersoff, in *Basic Concepts and Applications of Scanning Tunneling Microscopy and Related Techniques,* NATO Institute Series (Kluwer Academic, Norwell, MA, in press).

[12]The appearance of a hyperbolic sine here, in contrast to the simple exponential in Ref. 2, arises from the details of normalization, which are treated somewhat cavalierly here. The results are equivalent for large κR, and even in the worst case differ by only a factor of 2 in the coefficient, which causes very little change in the image. (For a discussion of the insensitivity of the image to even moderately large numerical changes in the coefficient, see for example Ref. 11).

[13]J. Tersoff, Phys. Rev. B **39**, 1052 (1989).

[14]J. Tersoff, Phys. Rev. Lett. **57**, 440 (1986).

[15]W. Sacks, S. Gauthier, S. Rousset, and J. Klein, Phys. Rev. B **37**, 4489 (1988).

[16]For air-exposed surface and tip, very large corrugations are typically measured, apparently due to contamination-mediated mechanical deformation of the surface or tip. See H. J. Mamin, E. Ganz, D. W. Abraham, R. E. Thomson, and J. Clarke, Phys. Rev. B **34**, 9015 (1986), and references therein. Under clean UHV conditions, R. H. Hamers found corrugations of 0.3 Å (private communication). Binnig *et al.* saw corrugations of roughly 1 Å [Europhys. Lett. **1**, 31 (1986)], but the fact that they occasionally saw even larger corrugations suggests that there may still have been some residual contamination problem. Of course, there is also some dependence of the image on voltage and, in principle, on current.